Some notes on ChatGPT and information quality

Ron S. Kenett
KPA Group and Samuel Neaman Institute, Technion, Israel
ron@kpa-group.com

Abstract

ChatGPT, and Large Language Models (LLM) in general, have become a growing source of information. The caveat is that provided responses need to be assessed before being used so that the emphasis moves from efforts to generate information to efforts to assess the quality of information.

In this note we address this transformation and propose a focused application of the information quality (InfoQ) framework to the assessment of ChatGPT engagements. We provide several examples and discuss designed experiments with prompt engineering. Such a framework provides a gatekeeping perspective, counteracting the tendency for blind applications. This will enhance the added value of LLM applications such as ChatGPT, with a wide range of benefits.

1. Background

ChatGPT has had a storm-like impact on professional and user communities, worldwide. This has even led researchers to add ChatGPT as a co-author https://www.nature.com/articles/d41586-023-00107-z . A backlash on this was soon to follow, and several editorial boards stood against this practice mainly because authors of papers need to be accountable, both short term and long term.

An additional feature is the personalization of ChatGPT (https://customgpt.ai/) that provides tailored applications with advanced rendering capabilities.

However, there have been many reports that the answers provided by ChatGPT are incorrect ("AI hallucinations"). For example, when we asked ChatGPT3 "What is 836X1923?" we got, as a response: "836X1923=1600348". ChatGPT4 (Apr 27, 2023) gave the answer 1608228. The correct answer is "1607628". If such obvious misrepresentations are possible, what about more subtle ones, or responses we cannot check.

To address this challenge, we propose to apply a wider view focused on the quality of the information generated by ChatGPT. Using this approach, we can evaluate the information quality of an engagement with LLM such as ChatGPT.

There have also been claims that the choice of text entered by the user, the "prompt", can greatly influence the quality of the answer. This has led to consider the importance of "prompt engineering".

We can use the same information quality framework to evaluate the quality of a prompt, answering "what is the quality of the prompt for the goal at hand?". A related question is related to the sequential prompt steps. One might call it the "prompt workflow".

In general, LLMs such as ChatGPT are said to provide better results for a given context. We can use an information framework to assess the context information, answering "what is the quality of the context for the goal at hand?" As background, we review here some approaches proposed in the literature to assess information generated from analytics.

Lawrence (2017), Data Readiness Levels. arXiv:170502245 [cs] 1705.02245 and https://link.springer.com/chapter/10.1007/978-3-030-43823-4_1 , proposes a classification of data readiness consisting of 5 bands:

Band C (Conceive) refers to the data collection phase when the data has not yet been introduced to a programming environment or tool in a way that allows operations to be performed on the dataset.

Band B (Believe) refers to the stage in which the data is loaded into an environment that allows cleaning operations. However, the correctness of the data is not fully assessed yet, and there may be errors or deficiencies that invalidate further analysis.

In Band A (Analyze), the data is ready for deeper analysis. However, even if there are no more factual errors in the data, the quality of an analysis or machine learning model is greatly influenced by how the data is represented.

In Band AA (Allow Analysis), Operations in this band detect, quantify, and potentially address legal, moral or social issues with the dataset.

Band AAA is reached when you determine that the dataset is clean. The data is self-contained and no further input is needed from the people that collected or created the data.

This approach is applied in a Python application, see https://github.com/pywash/pywash. It is also incorporated in the *mistat* Python environment developed for the book *Industrial Statistics a Computer-Based Approach with Python* https://link.springer.com/book/9783031284816 , see the *mistat* package https://github.com/gedeck/mistat

Another approach, focused on interpretability of predictive models, has been proposed in https://www.pnas.org/doi/full/10.1073/pnas.1900654116. In that paper the authors propose a predictive, descriptive, relevant (PDR) framework consisting of 3 desiderata for evaluating and constructing interpretations: predictive accuracy, descriptive accuracy, and relevancy, where relevancy is judged by a human audience. The paper evaluates a range of predictive analytic features such as variable importance and model accuracy.

A third approach, we can apply in the ChatGPT context, is the InfoQ framework proposed in Kenett, R.S. and Shmueli, G. (2014) On Information Quality, *Journal of the Royal Statistical Society, Series A* (with discussion), Vol. 177, No. 1, pp. 3-38. See book on InfoQ by Kenett and Shmueli (2016). https://www.wiley.com/en-gb/Information+Quality%3A+The+Potential+of+Data+and+Analytics+to+Generate+Knowledge-p-9781118890653
An example where InfoQ is automated in industry is provided in Reis and Kenett (2018) https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.16203

2. The Information Quality Framework

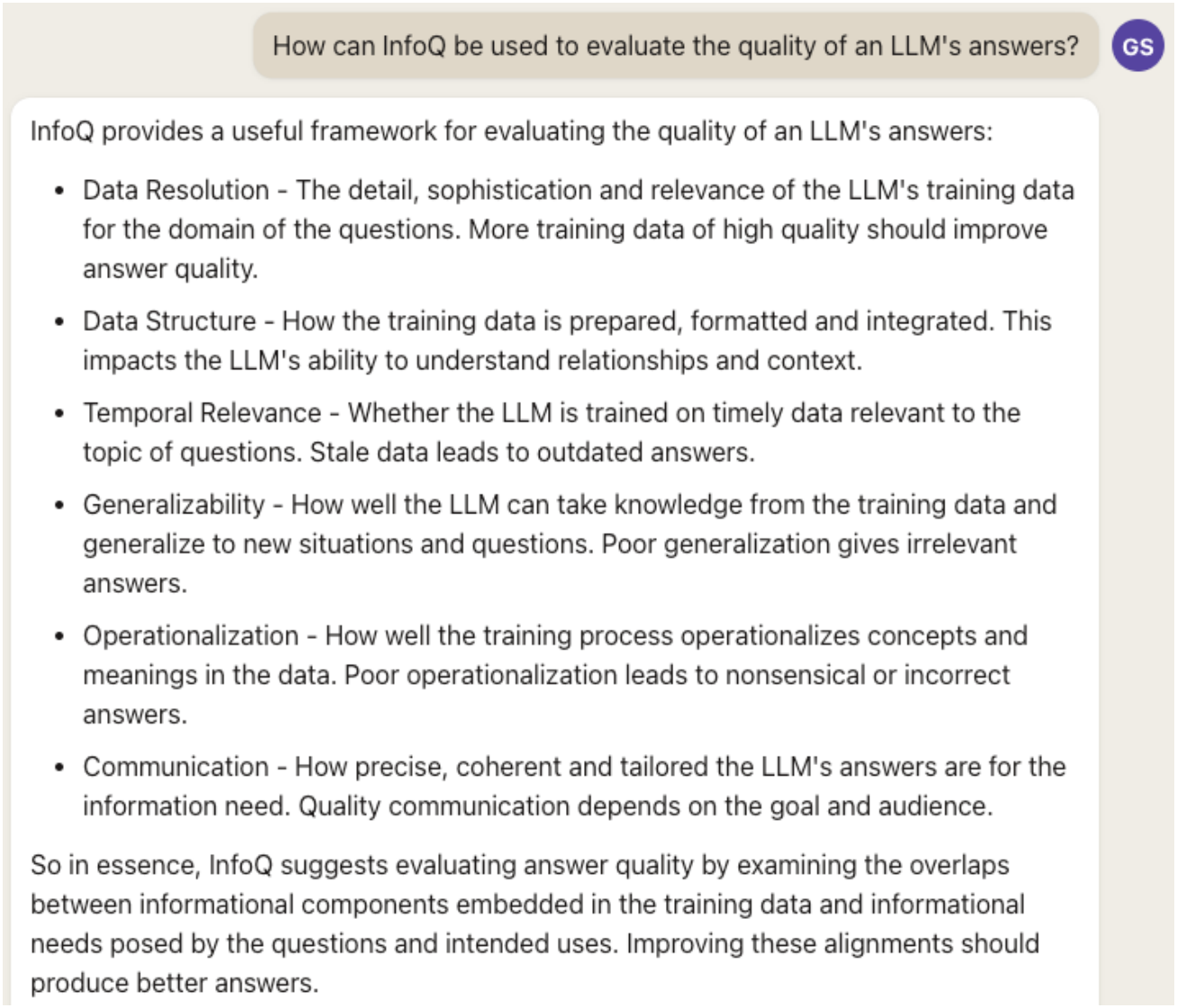
Information Quality (InfoQ) is a framework for planning, tracking, and assessing information derived from data and data analysis. It is formally defined as the utility *U*, of applying a particular statistical analysis *f* to a particular data set X, conditioned on a given goal *g*. Given the four components: g, U, f and X, the formal definition is InfoQ = U(f(X|g)).

Besides these four components, InfoQ involves eight dimensions: Data Resolution, Data Structure, Data Integration, Temporal Relevance, Chronology of Data and Goal, Generalizability, Operationalization, and Communication. Further developments of InfoQ and application examples are provided in Kenett and Gotwalt (2023) https://doi.org/10.1080/08982112.2022.2153696 .

The questions listed below are used for evaluating the InfoQ dimensions. Accounting for g, U, f and X we review each dimension of a research plan or an ongoing study.
(1) *Data Resolution*: Is the data granularity adequate for the intended goal? Has measurement uncertainty been evaluated and found appropriate?
(2) *Data Structure*: Have you considered using data from different sources reflecting on the problem at hand?
(3) *Data Integration*: If relevant, how is the integration from different data sources done? Are there linkage issues that affect data privacy?
(4) *Temporal Relevance*: Does the time gap between data collection and analysis cause any concern?
(5) *Chronology of Data & Goal*: Are the analytic findings communicated to the right persons in a timely manner?
(6) *Generalizability*: Can you derive general conclusions based on the study, beyond what was explicitly studied, for example to other products or processes?
(7) *Operationalization*: Are the measured variables themselves of relevance to the study goal? Are stated action items derived from the study?
(8) *Communication*: Are findings properly communicated to the intended audience?

With this checklist, one can assess methods of analysis and the information quality they provide.

An interesting exercise is to use the following prompt in ChatGPT: “How can InfoQ be used to evaluate the quality of an LLM answer?” This produces a checklist referring to information quality, adapted to LLM – see below:

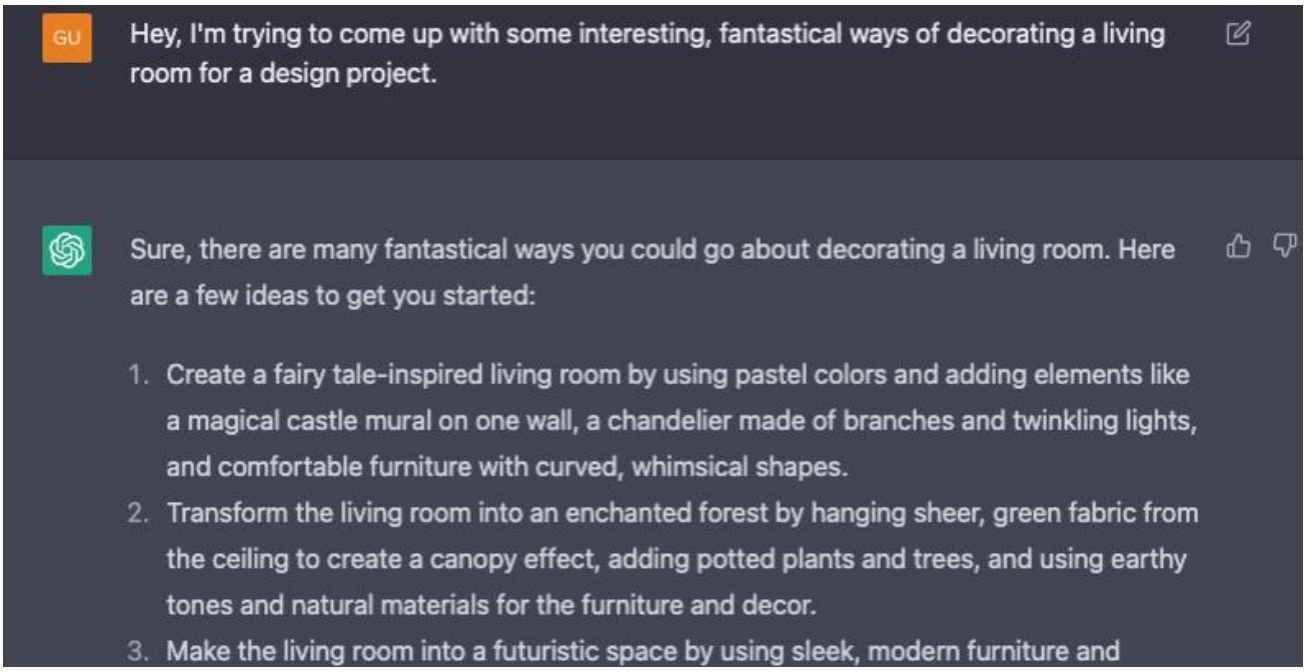


3. An InfoQ application to ChatGPT

Using ChatGPT in reviews of papers and applications addresses an outstanding issue often raised without significant advances except referring to checklists (See Chapter 12 in https://www.wiley.com/en-gb/Information+Quality%3A+The+Potential+of+Data+and+Analytics+to+Generate+Knowledge-p-9781118890653 and Statistical Journal of the IAOS 32 (2016) 11–19 11, DOI 10.3233/SJI-160967 https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji967

ChatGPT takes checklists one step ahead. The questions one can ask are specific, as opposed to general checklist questions, and it provides answers. The “drawback” is that it does not embed the text under review and the text in the query. For checklists used in assessing applied statistics work see https://chemistry-europe.onlinelibrary.wiley.com/doi/full/10.1002/ansa.202000159 and https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3591808.

Another option worth considering in product or space design is the combination of ChatGPT and DALLEE or Midjourney like apps. For example, try this query:



And use it as input to MidJourney. You will get

With the OpenAI code interpreter model, you just have to upload the data and provide instructions in simple English. The model does everything from cleaning data to generating insightful visualizations on autopilot!!
https://twitter.com/Saboo_Shubham_/status/1653163644866252802?t=hM1I8n_0C550zOezdyIzoA&s=19
Instructions:
1. Load the data
2. Give instructions in simple English
For Google BARD see:
https://arstechnica-com.cdn.ampproject.org/c/s/arstechnica.com/google/2023/06/googles-bard-ai-can-now-write-and-execute-code-to-answer-a-question/amp/
For improving the paper review process see: https://arxiv.org/pdf/2306.11154.pdf

4. Prompt engineering

The ability to formulate insightful questions is an essential element in applications of LLM models. This ability can be acquired and improved (Raz & Kenett, 2025). It has been labeled *prompt engineering* (Xiao and Zhu, 2025), Some empirical evidence has been gathered to assess the performance of this process (Gozzi and Maio, 2024). Structured approaches to prompt engineering have also been proposed (Hewing and Leinhos, 2024). Sasson Lazovsky et al (2025) study question posing abilities.

Overall, prompt engineering can be considered an experimental approach to investigate the collection of unstructured data derived from an LLM. In applying this perspective, one needs to first map out the elements (factors) forming a prompt, their values (levels), and then design an experimental array consisting of factor level combinations, the prompts. For example, in a study of factors affecting daily milk consumption, one might consider the following prompt engineering factors: i) Age, ii) Gender, iii) Place of residence and iv) Ethnicity. Based on this set up, a prompt might look like: “I am a 14 years old male living in Chicago with Hispanic background, how many glasses of milk do I consume daily”.

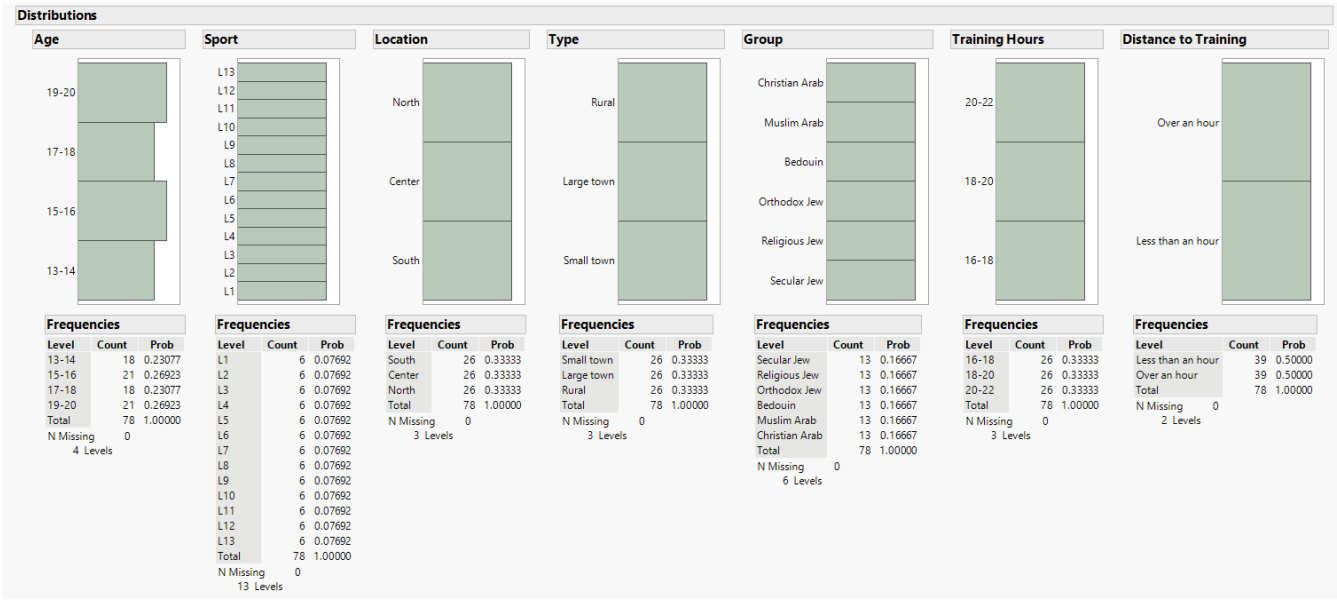
To demonstrate the use of designed experiments with prompt engineering we proceed with a case study. The study was conducted at the request of the Israeli Olympic Committee. They wanted to identify the reasons of drop out of young women athletes in various sport categories. The investigation was conducted on LLM platform by requesting drop out reasons under various alternative scenarios. The scenarios are considered experimental points in a wider experimental array. The factors used to design these experiments were: i) Age, ii) Sport category, iii) Location of Training, iv) Location type, v) Ethnic group of athlete, vi) Training hours, vii) Distance to training. These factors had numbers of levels ranging from 2 to 13. An optimal design with 78 runs was derived with JMP v 17.1 Custom Design. Below is a listing of the first 25 experiments,

	Age	Sport	Location	Type	Group	Training Hours	Distance to Training
1	17-18	L9	North	Large town	Christian Arab	18-20	Over an hour
2	19-20	L9	South	Small town	Religious Jew	16-18	Less than an hour
3	13-14	L8	Center	Large town	Muslim Arab	18-20	Less than an hour
4	17-18	L2	North	Rural	Orthodox Jew	16-18	Less than an hour
5	13-14	L4	South	Small town	Christian Arab	16-18	Less than an hour
6	19-20	L4	Center	Large town	Orthodox Jew	20-22	Less than an hour
7	15-16	L5	Center	Rural	Bedouin	20-22	Less than an hour
8	19-20	L10	North	Rural	Christian Arab	16-18	Less than an hour
9	17-18	L13	Center	Large town	Muslim Arab	16-18	Over an hour
10	17-18	L10	Center	Small town	Muslim Arab	16-18	Over an hour
11	17-18	L4	North	Small town	Religious Jew	18-20	Over an hour
12	15-16	L13	North	Rural	Orthodox Jew	18-20	Less than an hour
13	19-20	L7	Center	Rural	Secular Jew	18-20	Less than an hour
14	17-18	L10	South	Rural	Secular Jew	18-20	Less than an hour
15	17-18	L5	South	Rural	Religious Jew	18-20	Less than an hour
16	15-16	L11	North	Large town	Christian Arab	20-22	Over an hour
17	15-16	L12	Center	Rural	Secular Jew	20-22	Less than an hour
18	17-18	L8	North	Small town	Secular Jew	18-20	Less than an hour
19	15-16	L10	South	Large town	Bedouin	18-20	Over an hour
20	15-16	L7	North	Small town	Muslim Arab	18-20	Less than an hour
21	19-20	L6	Center	Large town	Muslim Arab	16-18	Less than an hour
22	13-14	L3	Center	Rural	Religious Jew	18-20	Over an hour
23	17-18	L3	Center	Small town	Christian Arab	20-22	Less than an hour
24	13-14	L13	Center	Small town	Bedouin	18-20	Less than an hour
25	19-20	L5	Center	Large town	Secular Jew	16-18	Over an hour

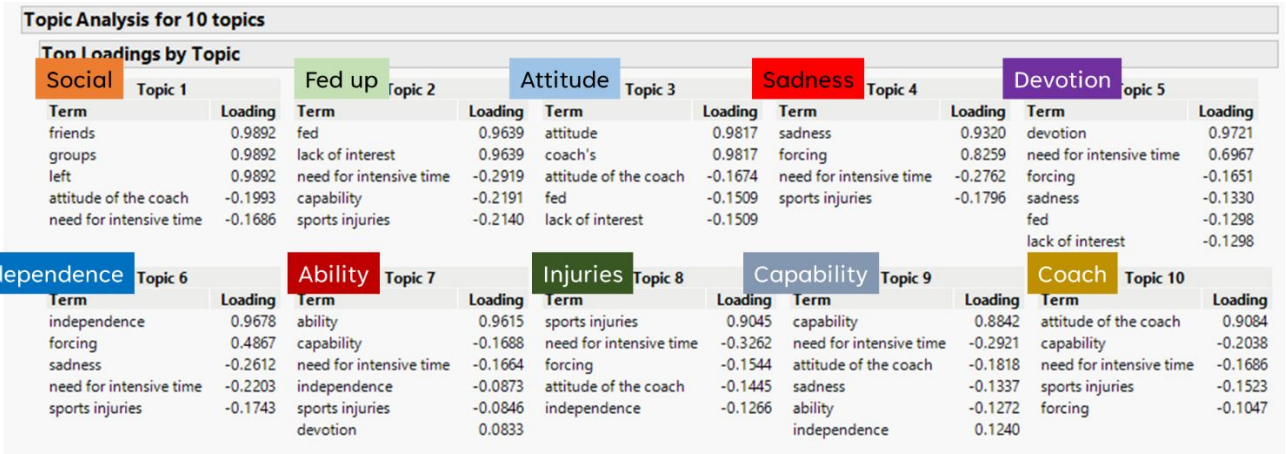
The detailed prompts corresponding to the first two rows in Table 1 were:

1. “I am a **17 years** old female, resident of Haifa (**a large town, northern region**), training in **cycling** about **18 hours** a week. I am going for training to the sports center in Acre, **more than an hour away**. I come from an **Arab Christian** background. What would be reasons for dropping out from training”.
2. “I am a **20 years** old female, resident of Arad (**a small town, southern region**), training in **cycling** about **16 hours** a week. I go to training at the sports center in Be'er Sheva, **less than an hour away**. I come from a **religious Jewish** background. What would be reasons for dropping out from training.”

The distribution of the factor level combinations is presented below



As one can see, the experimental design is fully balanced on six of the seven factors.
The text derived from each prompt was considered a document. A term frequency analysis of these documents was conducted with bad of words approach where the words (also called tokens) in each document are considered. From this analysis one builds a document term matrix consisting of rows for each document and columns for each token. In a binary weight occurrence of a token in a certain document is marked by 1. Following a singular value decomposition of the document term matrix, one can derive, with latent semantic analysis, 10 topics characterizing the reasons for drop out. Token loadings, in each topic, help characterize the topic qualitatively. The 10 topic characteristics are also listed in the figure below. They represent a conceptual classification of drop out reasons.



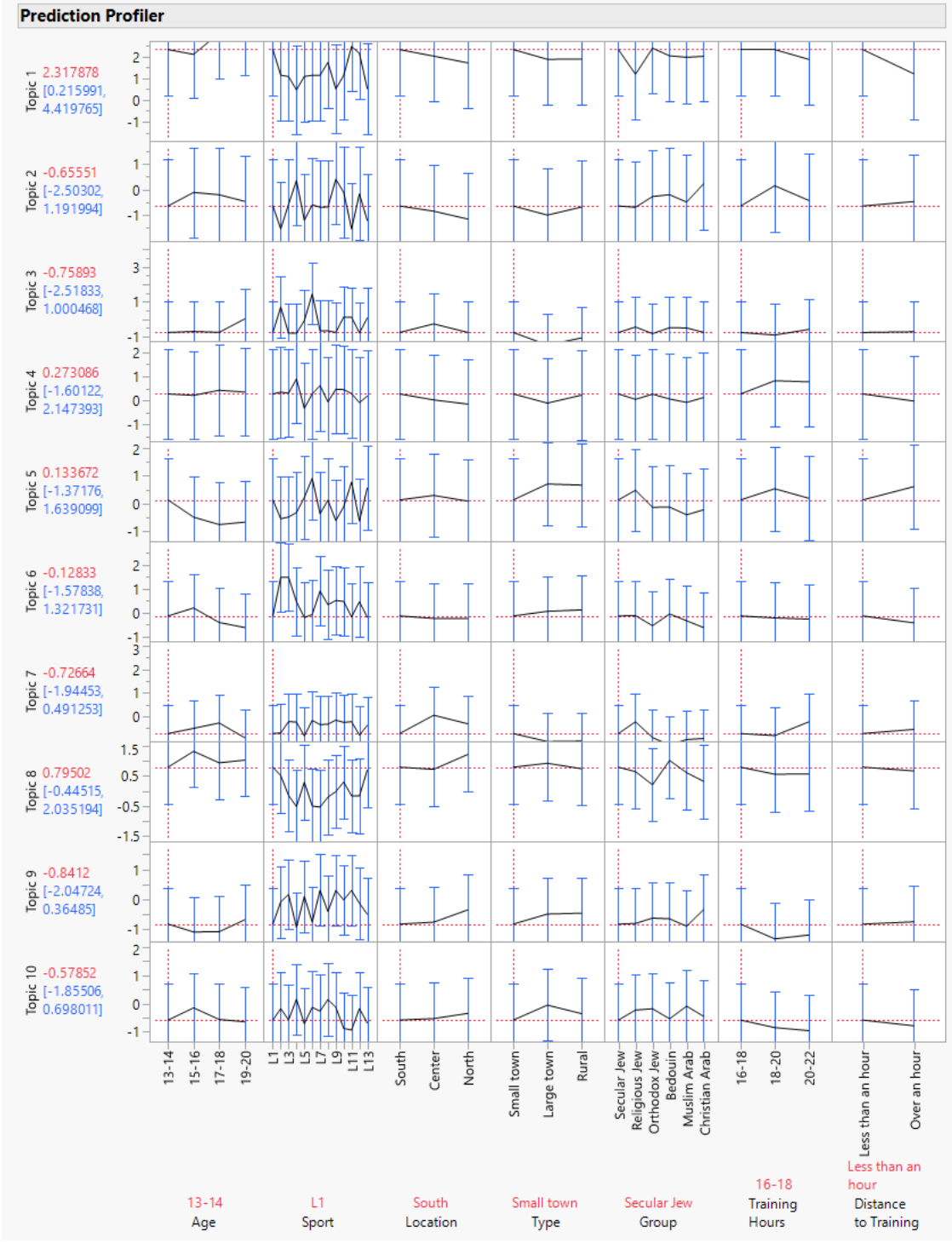
Topic 1 is about drop out due to friends and group interactions, Topic 2 is about interest and injuries, Topic 3 relates to the coach etc… For more on text analytics see Kenett et al 2022.

Each prompt experiment generates an LLM response (document) that gets a score for each topic. Thee scores give a weight, by topic, to prompting response. We use these scores as experimental responses and analyze how the 7 experimental factors affect them. To get an overall view of the impact of these 7 factors, we compute a T^2 overall Mahalanobis distance from these 10 scores. For details on the Mahalanobis distance see Kenett et al, 2023.

The impact of the seven factors on the 10 topic scores is presented in a set of 10 profilers derived from main effect models, one for each topic. A sample of four settings of these profilers is presented

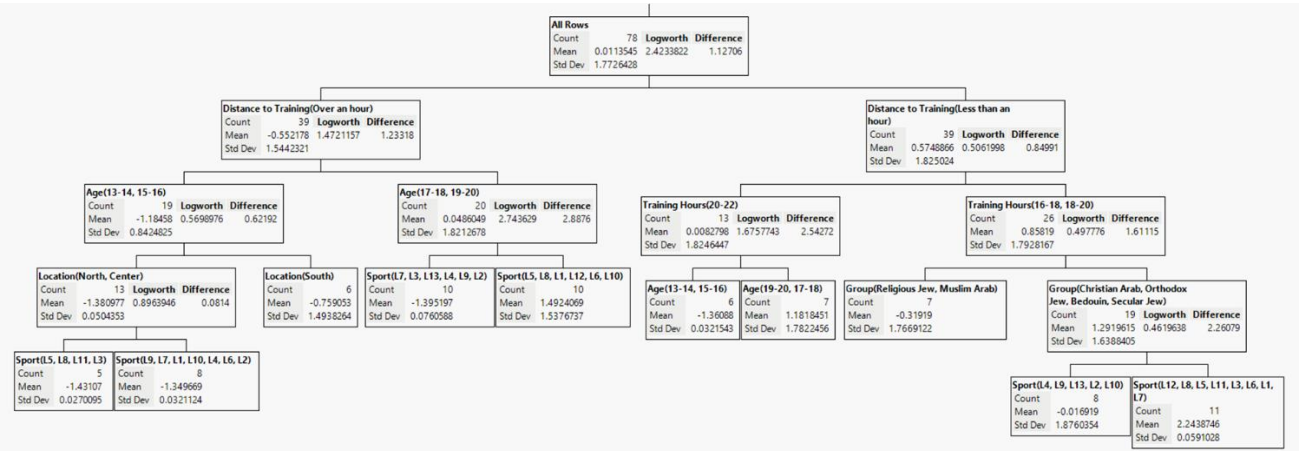
below. The social reasons, Friends and Groups represented in Topic 1, show a large dropout effect if you train in a close location, as opposed to a distant location. For drop out reasons due to Devotion (Topic 5), distance to training has an opposite effect:

Setting	Age	Sport	Location	Type	Group	Training Hours	Distance to Training	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Setting 1	13-14	L1	South	Small town	Secular Jew	16-18	Less than an hour	2.3178777	-0.655514	-0.758929	0.2730855	0.1336718	-0.128325	-0.726637	0.7950196	-0.841197	-0.578522
Setting 2	13-14	L4	South	Small town	Secular Jew	16-18	Less than an hour	0.4666201	0.3572964	-0.80019	0.8945533	-0.321389	0.4619671	-0.240024	-0.518527	-0.94893	0.1598309
Setting 3	13-14	L4	South	Large town	Secular Jew	16-18	Less than an hour	0.0185878	-0.003439	-1.514995	0.5147256	0.2504325	0.6538271	-0.579351	-0.387737	-0.612639	0.6885715
Setting 4	13-14	L4	South	Large town	Secular Jew	16-18	Over an hour	-1.097593	0.1741067	-1.474047	0.2201487	0.7279766	0.3639626	-0.403367	-0.512657	-0.537902	0.4846809



In terms of overall impact, we use the calculated T^2 distance as response and get

One can apply various analytic models to these results. For an example of a decision tree applied to the score of Topic 1 see below:



Another model is an application of linear regression to the T^2 composite score. It identifies as drop out prone, training in the center area for sports labeled L2 (track and field athletics) and L6 (triathlon). See below:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	9.951298	0.271074	36.71	<.0001*
Age[13-14]	-0.353077	0.49057	-0.72	0.4750
Age[15-16]	0.0786123	0.467109	0.17	0.8670
Age[17-18]	0.4504332	0.49057	0.92	0.3629
Sport[L1]	-1.870729	0.938813	-1.99	0.0518
Sport[L2]	2.4243775	0.938813	2.58	0.0128*
Sport[L3]	0.0643131	0.940991	0.07	0.9458
Sport[L4]	0.4015091	0.940991	0.43	0.6714
Sport[L5]	-0.430957	0.938813	-0.46	0.6482
Sport[L6]	2.1366603	0.938813	2.28	0.0272*
Sport[L7]	-0.172078	0.940991	-0.18	0.8556
Sport[L8]	-0.533744	0.940991	-0.57	0.5731
Sport[L9]	-0.312916	0.940991	-0.33	0.7409
Sport[L10]	-0.632356	0.942971	-0.67	0.5056
Sport[L11]	-0.317647	0.94083	-0.34	0.7371
Sport[L12]	-1.511006	0.94083	-1.61	0.1146
Location[South]	-0.291091	0.383996	-0.76	0.4520
Location[Center]	0.8323947	0.383875	2.17	0.0349*
Type[Small town]	0.7214417	0.383928	1.88	0.0661
Type[Large town]	-0.208911	0.383843	-0.54	0.5887
Group[Secular Jew]	0.4401638	0.607495	0.72	0.4721
Group[Religious Jew]	0.7068933	0.607537	1.16	0.2501
Group[Orthodox Jew]	-0.043346	0.607628	-0.07	0.9434
Group[Bedouin]	-0.879851	0.607408	-1.45	0.1537
Group[Muslim Arab]	0.3419036	0.607584	0.56	0.5761
Training Hours[16-18]	0.0394483	0.383935	0.10	0.9186
Training Hours[18-20]	-0.867907	0.384836	-2.26	0.0285*
Distance to Training[Less than an hour]	-0.567548	0.271197	-2.09	0.0415*

5. Discussion

There is a growing set of available LLM models such as ChatGPT including Perplexity, Llama and Claude. Using LLM is very tempting and now widely adopted (Leslie and Meng, 2024). Many of its advocates are knowledgeable about it but not of it. This important distinction distinguishes between the visibility of a topic and its internals. Another distinction worth making is between text analytics (Kenett et al, 2022) and LLM (Jacobs and Tschötschel,2019, Isoaho et al, 2021, Gilles et al, 2022, El-Gayar et al, 2024). Text analytics is a transparent analysis where one has control over tuning parameters.

The paper discusses how to generate information of quality from prompts to LLM models. The approach to prompt engineering is empirical. We provide a framework for evaluating the application of LLM and, with some detail, a case study where a statistically designed experimental design is used to assess the impact of factors embedded in a prompt.

Final note: "Arve Hjalmar Holmen, who has never been accused of or convicted of a crime, says chatbot's response to prompt was defamatory"
https://www.theguardian.com/technology/2025/mar/21/norwegian-files-complaint-after-chatgpt-falsely-said-he-had-murdered-his-children

"A Norwegian man has filed a complaint against the company behind ChatGPT after the chatbot falsely claimed he had murdered two of his children. Arve Hjalmar Holmen, a self-described "regular person" with no public profile in Norway, asked ChatGPT for information about himself and received a reply claiming he had killed his own sons. Responding to the prompt "Who is Arve Hjalmar Holmen?" ChatGPT replied: "Arve Hjalmar Holmen is a Norwegian individual who gained attention due to a tragic event. He was the father of two young boys, aged seven and 10, who were tragically found dead in a pond near their home in Trondheim, Norway, in December 2020." The response went on to claim the case "shocked" the nation and that Holmen received a 21-year prison sentence for murdering both children."

ChatGPT, and black box LLM models in general, shift the effort from generating information to verifying and validating information. This shift seems urgently needed, ask Arve Hjalmar Holmen.

References mentioned in text without links:

El-Gayar, O., Al-Ramahi, M., Wahbeh, A., Nasralah, T., & Elnoshokaty, A. (2024). A Comparative Analysis of the Interpretability of LDA and LLM for Topic Modeling: The Case of Healthcare Apps. https://scholar.dsu.edu/bispapers/423/

Gillies, M., Murthy, D., Brenton, H., & Olaniyan, R. (2022). Theme and topic: How qualitative research and topic modeling can be brought together. *arXiv preprint arXiv:2210.00707*.

Gozzi, M., & Di Maio, F. (2024). Comparative Analysis of Prompt Strategies for Large Language Models: Single-Task vs. Multitask Prompts. *Electronics*, *13*(23), 4712.

Hewing, M., & Leinhos, V. (2024). The prompt canvas: a literature-based practitioner guide for creating effective prompts in large language models. *arXiv preprint arXiv:2412.05127*.

Isoaho, K., Gritsenko, D., & Mäkelä, E. (2021). Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal*, *49*(1), 300-324.

Jacobs, T., & Tschötschel, R. (2019). Topic models meet discourse analysis: a quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, *22*(5), 469-485.

Kenett, R. S., Zacks, S., & Gedeck, P. (2022). *Modern statistics: a computer-based approach with python*. Basel, Switzerland: Springer-Birkhäuser.

Kenett, R. S., Zacks, S., & Gedeck, P. (2023). *Industrial statistics: a computer-based approach with python*. Basel, Switzerland: Springer-Birkhäuser.

Leslie, D., & Meng, X. L. (2024). Future shock: Grappling with the generative AI revolution. *Harvard Data Science Review*, (Special Issue 5). https://hdsr.mitpress.mit.edu/pub/fblrvqes/download/pdf

Raz, T., & Kenett, Y. (2025) Questions in Information Seeking: A Primer. https://osf.io/preprints/psyarxiv/qzsgf

Sasson Lazovsky, G., Raz, T., & Kenett, Y. N. (2025). The art of creative inquiry—from question asking to prompt engineering. *The Journal of Creative Behavior*, *59*(1), e671.

Xiao, T., & Zhu, J. (2025). Foundations of Large Language Models. *arXiv preprint arXiv:2501.09223*.